

Statistical inference for nested data using sufficient summary statistics: a tutorial

Irene Dowding^{a,*}, Stefan Haufe^{a,*}

^a*Technische Universität Berlin*

Abstract

Data with hierarchical structure arise in many fields. Estimating global effect sizes from nested data, and testing effects against global null hypotheses, is, however, more challenging than in the traditional setting of independent and identically distributed data. In this paper, we review statistical approaches to deal with nested data following either a fixed-effect or a random-effects model. We focus on methods that are easy to implement, such as group-level t-tests and Stouffer's method. The properties of these approaches are discussed within the context of neuroimaging applications, quantitatively assessed on simulated data, and demonstrated on real human neurophysiological data from a simulated-driving experiment. With what we call the inverse-variance-weighted sufficient-summary-statistic approach, we highlight a particularly compelling technique that combines desirable statistical properties with computational simplicity, and we provide step-by-step instructions to apply it to a number of popular measures of effect size.

Keywords: hierarchical inference, non-i.i.d. data, effect size, fixed effect, random effects, hypothesis testing, statistical power, summary statistic, inverse-variance-weighting, t-test, correlation, area under the curve, Stouffer's method, neuroimaging, electroencephalography

1. Introduction

In many applications of statistics, data possess a nested (hierarchical) structure. Such data arises naturally due to grouping of subjects (e.g., patients in different medical centers, students taught by different teachers, data acquired in different scientific studies) or due to repeated measurements in the same subject. In those situations, the assumption of identical distributed observations is not appropriate. In multi-center clinical trials, for example, subtle differences may exist in each of the centers' processes as well as in the genetic or cultural traits of the studied populations.

In psychological or neuroimaging studies, multiple data points are typically acquired for the same subject throughout the course of an experiment. While samples acquired from the same subject may be expected to be identically distributed, differences must be assumed across subjects. In specific application such as functional magnetic resonance imaging (fMRI), different distributions are typically even assumed for individual experimental sessions conducted with the same subject, giving rise to a three-level nesting hierarchy. Here we consider the setting of standard neuroimaging studies, and distinguish between a subject (lower) and a group (higher) level.

From a statistical point of view, the question arises how to obtain precise (e.g., best unbiased) estimators for group-level effect sizes as well as powerful statistical tests for nested data. A flexible statistical framework for such data consists of combining data of all subject in one single linear model, referred to as the nested linear model, hierarchical linear model,

multi-level model or linear mixed model. These models are extensively covered in, e.g., [38, 54, 23]. Estimating hierarchical linear models typically requires advanced statistical techniques that can be hard to implement and difficult to optimize. In some applications such as fMRI, for example, analyzing the combined data of all subjects within a single model poses a serious computational challenge. In other applications such as meta analyses, this may not even be possible as only summary statistics but no individual samples are available from each study.

We here consider the problem of estimating group-level effect sizes and estimating their statistical significance. We point out that this problem can be formulated in a compellingly simple framework, where group-level inference is conducted using the results of separate subject-level analyses only. The resulting statistical methods are simple to implement, computationally efficient to estimate, and the underlying model can be made equivalent to a nested linear model in many cases [2]. Such approaches are standard in meta analysis [4, 6], and known as 'summary statistic' approaches in the fMRI literature [22, 2, 31].

Within this framework, this paper reviews ways to estimate group-level effect sizes and to assess their statistical significance. We first provide a reference for a number of popular parametric and non-parametric effect size measures (Section 2.2). We then discuss the need to choose an appropriate group-level model, as between-subject variances typically have to be taken into account through a 'random effects' model as opposed to a 'fixed effect' model (Section 2.3). We also demonstrate why the simple approach of ignoring the group structure by pooling the data of all subject is invalid (Section 2.4). We then outline the popular approach of computing effect sizes on the subject level and treating these effect

*Authors contributed equally. Corresponding author.

Email addresses: irenedowding@web.de (Irene Dowding), stefan.haufe@tu-berlin.de (Stefan Haufe)

sizes as samples in a group-level test. This ‘naive-summary-statistic’ approach is, however, suboptimal in terms of statistical power (Section 2.5). With what we call the inverse-variance-weighted sufficient-summary-statistic approach, we discuss a technique that overcomes these limitations, leading to unbiased minimum-variance group-level effect size estimates and a powerful statistical test of group-level null hypotheses. In a tutorial style, we outline the steps that are required to employ this approach (Section 2.6). Lastly, we discuss the advantages and drawbacks of Stouffer’s method of combining subject-level p-values (Section 2.7).

Using synthetic data representing a two-sample separation problem, we empirically assess the sensitivity and specificity of the reviewed approaches (Section 3). The properties of the various approaches are further highlighted in an application to human neurophysiological data acquired during simulated emergency braking in a driving simulator (Section 4). The paper ends with a discussion of nested linear models, of multivariate extensions, of non-parametric (bootstrapping and surrogate data) approaches, and a note on prominent applications of nested statistical models in neuroimaging (Section 5).

2. Theory

2.1. Statistical terminology

An *effect size* θ is any quantitative measure that reflects the magnitude of some phenomenon of interest. An *estimator* $\hat{\theta}$ for θ is *unbiased*, if its expected value is θ .

A *statistical test* is a procedure to decide, based on an observed sample from a population, which of two complementary hypotheses about a population parameter is true. In this paper, we want to make inference about the presence or absence of an effect. The *null hypothesis* is that no effect is present. The zero effect is denoted by θ_0 . The null hypothesis is denoted by $H_0 : \theta = \theta_0$. The *alternative hypothesis* that an effect is present is denoted by H_1 . A *one-tailed* alternative hypothesis assumes that either $H_1 : \theta > \theta_0$ or $H_1 : \theta < \theta_0$, while a *two-tailed* alternative hypothesis assumes that $H_1 : \theta \neq \theta_0$.

The *p-value* denotes the probability of observing an effect at least as strong as the observed effect under the assumption of the null hypothesis. To calculate the p-value, a *test statistic* needs to be derived from the observed effect size, where its distribution under the null hypothesis is known or can be reasonably well approximated. Denoting the test statistic by T , its cumulative distribution function under the H_0 by F_T , and its observed value in a given sample by τ , the p-values for a *one-tailed* alternative hypothesis are given by

$$Pr(T \leq \tau | H_0) = F_T(\tau) \quad (1)$$

$$Pr(T \geq \tau | H_0) = 1 - F_T(\tau), \quad (2)$$

where $Pr(\cdot)$ denotes probability. The p-value for a two-tailed alternative hypothesis is given by

$$2 \cdot \min \{Pr(T \leq \tau | H_0), Pr(T \geq \tau | H_0)\}. \quad (3)$$

The null hypothesis is rejected (and the alternative hypothesis accepted) if the p-value falls below an *alpha-level* α . The most commonly used alpha-levels are $\alpha = 0.05$ and $\alpha = 0.01$.

If the null hypothesis is rejected, we speak of a *statistically significant* effect. The value of the test statistic that is required for a significant effect is called *critical value*.

The *power* or *sensitivity* of a statistical test is the fraction of actually present effects that are statistically significant according to the test. The *specificity* of a test is the fraction of non-existent effects, for which the H_0 is accepted. Conversely, the *false positive rate* is the fraction of non-existent effects that are statistically significant.

2.2. Common effect size measures

Before introducing the nested data setting, we review a number of popular effect size measures. For each measure, we also present an analytic expression of its variance, which is a prerequisite for assessing its statistical significance. We will later need the variance for performing statistical inference in the nested setting, too.

2.2.1. Mean of a sample

A common measure of effect size is the mean of a sample. Consider a neuroimaging experiment, in which the participant is repeatedly exposed to the same stimulus. A common question to ask is whether this stimulus evokes a brain response that is significantly different from a baseline value. Assume that we observe N independent samples $x_1, \dots, x_N \in \mathbb{R}$. The sample mean is denoted by $\bar{x} = 1/N \sum_{i=1}^N x_i$, and the unbiased sample variance is given by $\hat{\sigma}_x^2 = 1/(N-1) \sum_{i=1}^N (x_i - \bar{x})^2$. The variance of \bar{x} is given by

$$\widehat{\text{Var}}(\bar{x}) = \frac{\hat{\sigma}_x^2}{N}. \quad (4)$$

Assuming *independent and identically distributed* (i.i.d.) samples, which are either normal (Gaussian) distributed or large enough, the null hypothesis $H_0 : \bar{x} = \theta_0$ can be tested using that the statistic

$$t = \frac{\bar{x} - \theta_0}{\sqrt{\widehat{\text{Var}}(\bar{x})}} \quad (5)$$

is approximately Student-t-distributed with $N - 1$ degrees of freedom. This is the *one-sample t-test*.

A similar effect size is the mean difference $\overline{x - y} = 1/N \sum_{i=1}^N x_i - y_i$ of two paired samples $(x_1, y_1), \dots, (x_N, y_N) \in \mathbb{R}^2$. Here, the y_i could, for example, represent baseline activity that is measured in each repetition right before the presentation of the experimental stimulus. A natural null hypothesis is that the mean difference is zero, i.e., $H_0 : \overline{x - y} = 0$. This hypothesis can be tested with a *paired t-test*, which replaces x by $x - y$ in Eq. (4),(5).

Note that, if x or y cannot be assumed to be normal distributed, a more appropriate test is the non-parametric *Wilcoxon signed-ranked test* [52].

2.2.2. Difference between class-conditional means

A slightly different treatment is required for the difference between the means of two unpaired samples. Consider an experiment with two conditions \mathcal{X} and \mathcal{Y} . In neuroimaging studies, these could differ in the type of stimulus presented.

We observe N_X samples $x_1, \dots, x_{N_X} \in \mathbb{R}$ of brain activity within condition \mathcal{X} , and N_Y samples $y_1, \dots, y_{N_Y} \in \mathbb{R}$ within condition \mathcal{Y} . The sample means are denoted by \bar{x} and \bar{y} , and their difference is given by

$$\hat{d} = \bar{x} - \bar{y}. \quad (6)$$

The variance of \hat{d} is estimated as

$$\widehat{\text{Var}}(\hat{d}) = \frac{\hat{\sigma}_x^2}{N_X} + \frac{\hat{\sigma}_y^2}{N_Y}, \quad (7)$$

where $\hat{\sigma}_x^2$ and $\hat{\sigma}_y^2$ are the unbiased sample variances of \mathcal{X} and \mathcal{Y} . The null hypothesis of equal means is given by $H_0 : d = 0$. Under the assumption of either normal distributed x_i and y_i , or large enough samples, the null hypothesis can be tested with *Welch's two-sample t-test*. It computes the test statistic

$$t = \frac{\hat{d}}{\sqrt{\widehat{\text{Var}}(\hat{d})}} \quad (8)$$

which is approximately Student-t-distributed. The degrees of freedom can be approximated using the Welch-Satterthwaite equation [50]. Note that assuming equal variances of \mathcal{X} and \mathcal{Y} leads to the better known Student's t-test, which is, however, less recommendable than Welch's t-test [40].

2.2.3. Area under the ROC curve

In many cases, one may be interested in quantifying the predictive accuracy of a binary classifier to separate experimental condition \mathcal{X} from condition \mathcal{Y} . The *receiver operating characteristic (ROC)* is a plot that visualizes the performance of such a binary classification system. It is obtained by plotting the true positive rate (TPR) against the false positive rate (FPR) when varying the threshold that divides the predicted condition into \mathcal{X} and \mathcal{Y} . Assume without loss of generality that condition \mathcal{X} is associated with a positive label indicating that the detection of instances of that condition is of particular interest, while \mathcal{Y} is associated with a negative label. TPR is defined as the fraction of correctly classified positive samples among all positive samples, while FPR denotes the fraction negative samples that are incorrectly classified as positives.

A common way to reduce the ROC curve to a single quantity is to calculate the area beneath it [11]. The resulting statistics is called the *area under the curve (AUC)*, and is equivalent to the probability that a classifier will correctly rank a randomly chosen pair of samples (x, y) , where x is a sample from \mathcal{X} and y is a sample from \mathcal{Y} [17]. The AUC is also equivalent [see 17, 30] to the popular Mann-Whitney U [29] and Wilcoxon rank-sum [52] statistics, which provide a non-parametric test for differences in the central tendencies of two unpaired samples. It is therefore an appropriate alternative to the two-sample t-test discussed in Section 2.2.2, if the data follow non-Gaussian distributions.

Assuming, without loss of generality, that higher values are indicative for class \mathcal{X} , the AUC is given as

$$A = \Pr(x > y | x \in \mathcal{X}, y \in \mathcal{Y}). \quad (9)$$

Perfect class separability is denoted by $A = 0$ and $A = 1$, while chance-level class separability is attained at $A = 0.5$. Thus, a common null hypothesis is $H_0 : A = 0.5$.

Assume we have N_X samples from condition \mathcal{X} and N_Y samples from condition \mathcal{Y} . To compute the test statistics, all observations from both conditions are pooled and ranked, beginning with rank one for the smallest value. Defining by $\text{rank}(x_n)$ the rank of x_n (the n -th sample from condition \mathcal{X}), the Wilcoxon rank-sum statistic for class \mathcal{X} is defined as

$$W = \sum_{n=1}^{N_X} \text{rank}(x_n), \quad (10)$$

while the Mann-Whitney U statistic is given by

$$U = W - \frac{N_X(N_X + 1)}{2}. \quad (11)$$

Finally, the AUC statistic is given by

$$\hat{A} = \frac{U}{N_X N_Y}. \quad (12)$$

The exact distributions of W , U and \hat{A} under the null hypothesis can be derived from combinatorial considerations [29, 30], and critical values for rejecting the null hypothesis can be calculated using recursion formulae [43]. However, these distributions are approximately normal distributed for samples of moderate size ($N_X + N_Y \geq 20$). The mean and variance of Mann-Whitney's U is given by

$$\mathbb{E}_{H_0}(U) = \frac{N_X N_Y}{2} \quad \text{Var}_{H_0}(U) = \frac{N_X N_Y (N_X + N_Y + 1)}{12}, \quad (13)$$

where $\mathbb{E}_{H_0}(\cdot)$ and $\text{Var}_{H_0}(\cdot)$ denote expected value and variance under the null hypothesis [30]. From Eq. (12), the mean and variance of the AUC statistic follow as

$$\mathbb{E}_{H_0}(\hat{A}) = \frac{1}{2} \quad \text{Var}_{H_0}(\hat{A}) = \frac{\text{Var}_{H_0}(U)}{N_X^2 N_Y^2}. \quad (14)$$

Note that this null distribution does not depend on the distribution of the data, and is only based on the assumptions of i.i.d. samples, equal variances of both classes, and that observations are ordinal (that is, it is possible to rank any two observations).

If the null hypothesis is violated (e.g., $A \neq 0.5$), the variances of U , W , and \hat{A} become data-dependent. The variance for general A can be approximated as [17, 15]

$$\widehat{\text{Var}}(\hat{A}) = \frac{\hat{A}(1 - \hat{A}) + (N_X - 1)(Q_1 - \hat{A}^2) + (N_Y - 1)(Q_2 - \hat{A}^2)}{N_X N_Y}, \quad (15)$$

where $Q_1 = \hat{A}/(2 - \hat{A})$ and $Q_2 = (2\hat{A}^2)/(1 + \hat{A})$. The variances of U and W follow accordingly. A statistical test for the null hypothesis can be devised using that

$$z = \frac{\hat{A} - 0.5}{\sqrt{\widehat{\text{Var}}(\hat{A})}} \quad (16)$$

is approximately standard normal distributed for large sample sizes (analogous for U and W).

2.2.4. Pearson correlation coefficient

The *Pearson product-moment correlation coefficient* $\hat{\rho}$ is used when one is interested in the linear dependence of a pair of random variables (X, Y) . Suppose that for each subject, we have N i.i.d. pairs of observations $(x_1, y_1), \dots, (x_N, y_N) \in \mathbb{R}^2$ with sample mean (\bar{x}, \bar{y}) . In a neuroimaging context, these pairs could reflect neural activity in two different anatomical structures, or concurrently-acquired neural activity and behavioral (e.g. response time relative to a stimulus) data. The sample Pearson product-moment correlation coefficient is given by

$$\hat{\rho} = \frac{\sum_{n=1}^N (x_n - \bar{x})(y_n - \bar{y})}{\sqrt{\sum_{n=1}^N (x_n - \bar{x})^2} \sqrt{\sum_{n=1}^N (y_n - \bar{y})^2}}, \quad (17)$$

where $\hat{\rho} = 1$ denotes perfect correlation, and $\hat{\rho} = -1$ denotes perfect anti-correlation. The null hypothesis of no correlation is given by $H_0 : \rho = 0$. Assessing the statistical significance of Pearson correlations can be done using the *Fisher z-transformation* [13], defined as

$$\zeta(\hat{\rho}) := \frac{1}{2} \ln \left(\frac{1 + \hat{\rho}}{1 - \hat{\rho}} \right) = \operatorname{arctanh}(\hat{\rho}). \quad (18)$$

If (X, Y) has a bivariate normal distribution, then $\zeta(\hat{\rho})$ is approximately normal distributed with mean $\operatorname{arctanh}(\rho)$ and variance

$$\operatorname{Var}(\zeta(\hat{\rho})) = \frac{1}{N - 3}. \quad (19)$$

Therefore the test statistic

$$z = \frac{\zeta(\hat{\rho})}{\sqrt{\operatorname{Var}(\zeta(\hat{\rho}))}} \quad (20)$$

is approximately standard normal distributed.

The Fisher-transformation is also used when averaging correlations, where the standard approach is to Fisher-transform each individual correlation before computing the average. The reason behind this step is that the distribution of the sample correlation is skewed, whereas the Fisher-transformed sample correlation is approximately normal distributed and thus symmetric (cf. [44]). Results can be transformed back into a valid Pearson correlation using the inverse transformation

$$\hat{\rho} := \frac{e^{2\zeta(\hat{\rho})} - 1}{e^{2\zeta(\hat{\rho})} + 1} = \tanh(\zeta(\hat{\rho})). \quad (21)$$

The same back transformation can be applied to map confidence intervals derived for $\zeta(\hat{\rho})$ into the Pearson correlation domain.

Pearson correlation can also be used to derive the *coefficient of determination*, which indicates the proportion of the variance in the dependent variable that is predictable from the independent variable in a linear regression. If an intercept term is included in the regression, the coefficient of determination \hat{R} is given as the square of the Pearson product-moment correlation between the two variables $\hat{R} = \hat{\rho}^2$. Another strongly related quantity is the point-biserial correlation

coefficient, which is used when one variable is dichotomous, i.e., indicates membership in one of two experimental conditions. Pearson correlation is mathematically equivalent to point-biserial correlation if one assigns two distinct numerical values to the dichotomous variable.

2.2.5. Linear regression coefficients

A multiple linear regression model has the form

$$y_n = \beta_0 + x_{n,1}\beta_1 + \dots + x_{n,K}\beta_K + \eta_n, \quad (22)$$

where the dependent variable $y_n, n \in \{1, \dots, N\}$ is the n -th sample, $x_{n,k}, k \in \{1, \dots, K\}$ are independent variables (or, factors), β_1, \dots, β_K are corresponding regression coefficients, β_0 is an intercept parameter, and η_n is zero-mean, uncorrelated noise. In a neuroimaging context, the samples y_n could represent a neural feature such as the activity of a particular brain location measured at various times n , while the $x_{n,k}$ could represent multiple factors thought to collectively explain the variability of y_n such as the type of experimental stimulus or behavioral variables. In some fields, such a model is called a neural encoding model. It is also conceivable to have the reverse situation, in which the $x_{n,k}$ represent multiple neural features, while the dependent variable y_n is of non-neural origin. This situation would be called neural decoding.

The independent variables $x_{n,k}$ could be either categorical (i.e., multiple binary variables coding for different experimental factors) or continuous. The specific case in which all independent variables are categorical is called analysis of variance (ANOVA). Linear models therefore generalize a relatively broad class of effect size measures including differences between class-conditional means and linear correlations [37].

The most common way to estimate the regression coefficients $\beta_k, k \in \{0, \dots, K\}$ is ordinary least-squares (OLS) regression. The resulting estimate is also the maximum-likelihood estimate under the assumption of Gaussian-distributed noise. Using the vector/matrix notations $\mathbf{y} = (y_1, \dots, y_N)^\top$, $\boldsymbol{\beta} = (\beta_0, \dots, \beta_K)^\top$, $\boldsymbol{\eta} = (\eta_1, \dots, \eta_N)^\top$, $\mathbf{x}_n = (1, x_{n,1}, \dots, x_{n,K})^\top$, and $\mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_N]^\top \in \mathbb{R}^{N \times (K+1)}$, Eq. (48) can be rewritten as $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\eta}$. The OLS estimate is then given by

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}. \quad (23)$$

The estimated coefficients $\hat{\beta}_k$ can be treated as effect sizes measuring how much of measured data is explained by the individual factors $x_{n,k}$. The null hypothesis for factor k having no explanatory power is $H_0 : \beta_k = 0$. The estimated variance of $\hat{\beta}_k$ is

$$\widehat{\operatorname{Var}}(\hat{\beta}_k) = C_{k,k}, \quad (24)$$

where $C = \hat{\sigma}_\eta^2 (\mathbf{X}^\top \mathbf{X})^{-1}$ and $\hat{\sigma}_\eta^2 = \frac{1}{N - (K+1)} \sum_{n=1}^N (y_n - \hat{\boldsymbol{\beta}}^\top \mathbf{x}_n)^2$ is an unbiased estimator of the noise variance. A statistical test for the null hypothesis can be devised using that

$$t = \frac{\hat{\beta}_k}{\sqrt{\widehat{\operatorname{Var}}(\hat{\beta}_k)}} \quad (25)$$

is t-distributed with $N - (K + 1)$ degrees of freedom. A similar procedure can be devised for regularized variants such as Ridge regression [21].

2.3. Nested statistical inference

In the following, our goal is to combine the data of several subjects to estimate a population effect and to assess its statistical significance. We denote the number of subjects with S . The observed effect sizes of each individual subject are denoted by $\hat{\theta}_s$ ($s = 1, \dots, S$). Other quantities related to subject s are also indexed by the subscript s , while the same quantities without subject index denote corresponding group-level statistics.

Two different models may be formulated for the overall population effect.

Fixed-effect (FE) model. In the fixed-effect (FE) model, we assume that there is one (fixed) effect size θ that underlies each subject, that is

$$\theta_1 = \theta_2 = \dots = \theta_S =: \theta. \quad (26)$$

The observed effect $\hat{\theta}_s$ can therefore be modeled as

$$\hat{\theta}_s = \theta + \epsilon_s, \quad \text{with } \text{Var}(\epsilon_s) = \sigma_s^2, \quad (27)$$

where ϵ_s denotes the deviation of the subject's observed effect from the true effect $\theta_s = \theta$. We assume that the noise terms ϵ_s are independent, zero-mean random variables with subject-specific variance σ_s^2 .

The null hypothesis tested by a fixed-effect model is that no effect is present in any of the subjects. Thus, $H_0 : \theta = \theta_1 = \dots = \theta_S = \theta_0$, where θ_0 denotes the zero effect.

Random-effects (RE) model. In the random-effects (RE) model, the true effect sizes are allowed to vary over subjects. They are assumed to follow a common distribution of effects with mean θ . The observed effect $\hat{\theta}_s$ is modeled as

$$\theta_s = \theta + \xi_s \quad \text{with } \text{Var}(\xi_s) = \sigma_{\text{rand}}^2 \quad (28)$$

$$\hat{\theta}_s = \theta_s + \epsilon_s \quad \text{with } \text{Var}(\epsilon_s) = \sigma_s^2, \quad (29)$$

where ϵ_s denotes the deviation of the subject's observed effect from the true subject-specific effect θ_s , and where ξ_s denotes the deviation of the true subject-specific effect θ_s from the population effect θ . ξ_s and ϵ_s are assumed to be zero-mean, independent quantities. The subject-specific variance of ϵ_s is σ_s^2 , while the variance of ξ_s is σ_{rand}^2 . For $\sigma_{\text{rand}}^2 = 0$, we recover the fixed-effect model.

The null hypothesis being tested is that the population effect is zero ($H_0 : \theta = \theta_0$), while each individual subject-specific effect θ_s may still be non-zero.

Besides testing different null hypotheses, fixed-effect and random-effects models assume different variances of the observed effect sizes. In the fixed-effect model, all observed variability is assumed to be within-subject variability

$$\text{Var}(\hat{\theta}_s) = \sigma_s^2. \quad (30)$$

The random-effects model additionally accounts for variability between subjects

$$\text{Var}(\hat{\theta}_s) = \sigma_s^2 + \sigma_{\text{rand}}^2. \quad (31)$$

If the data follow a random-effects model, neglecting σ_{rand}^2 in a fixed-effect analysis leads to an underestimation of the variance. This has negative consequences if we attempt to make inference on the mean population effect ($H_0 : \theta = \theta_0$) relying only on a fixed-effect analysis: We may arrive at spurious results, as the underestimated variance leads to a decreased specificity of the statistical test [24, 12, 41]. On the other hand, there is little disadvantage of using a random-effects analysis, even when the data follows a fixed-effect model. As the assumption of a fixed population effect is unrealistic in most practical cases, it is often recommended to carry out random-effects analysis per default [12, 35, 6, 31].

2.4. Data pooling

The most naive approach to conduct group-level inference would be to pool the samples of all subjects, and thus to disregard the nested structure of the data. In electroencephalography (EEG) studies, this approach is sometimes pursued when computing 'grand-average' (group-level) waveforms of event-related potentials (ERP) that are elicited by the brain in response to a stimulus.

Pooling the samples of all subjects may violate the assumption of identically distributed data underlying many statistical tests. Depending on the type of analysis, this may result in an over- or underestimation of the effect size, an over- or underestimation of the effect variance, and ultimately in low sensitivity or low specificity of the resulting statistical test.

The following two examples illustrate the problem. In both cases, two variables, X and Y , are modeled for $S = 4$ subjects. $N = 20$ samples were independently drawn for each subject and variable from Gaussian distributions according to $x_{n,s} \sim \mathcal{N}(\mu_s - 1, 4)$, $y_{n,s} \sim \mathcal{N}(\mu_s + 1, 4)$, $s = 1, \dots, S$, $n = 1, \dots, N$, where the notation $\mathcal{N}(\mu, \sigma^2)$ denotes a Gaussian distribution with mean μ and variance σ^2 . The subject-specific offsets μ_s were independently drawn from another Gaussian: $\mu_s \sim \mathcal{N}(0, 15^2)$. In a practical example (e.g., neuroimaging), these means may indicate individual activation baselines, which are usually not of interest. Given the generated sample, a difference in the means of X and Y is correctly identified for each subject by Welch's two-sample t-test ($p \leq 0.02$). Because of the substantial between-subject variance, this difference is, however, not significant in the pooled data of all subjects ($p = 0.29$). See Figure 1 (A) for a graphical depiction.

A Pearson correlation analysis of the same data correctly rejects the hypothesis of a linear dependence between X and Y for each subject ($|r| \leq 0.14$, $p \geq 0.55$). However, the presence of subject-specific offsets μ_s causes a strong correlation of X and Y across the pooled data of all subjects ($r = 0.98$, $p \leq 10^{-16}$, see Figure 1 (B) for a depiction). In many practical cases, this correlation will not be of interest and must be considered spurious.

These examples motivate the use of hierarchical approaches for testing data with nested structure, which we introduce below.

2.5. Naive summary-statistic approach

The simplest variant of the summary-statistic approach ignores subject-specific variances σ_s^2 , treating subject-level ef-

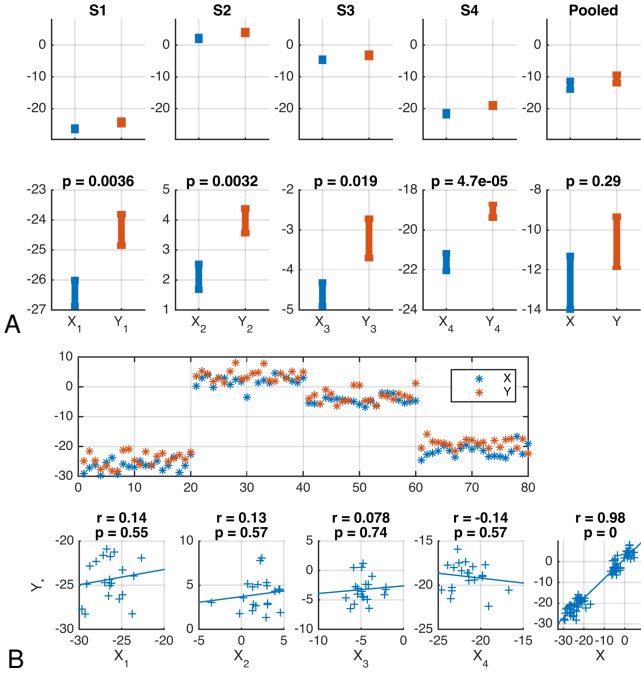


Figure 1: Low sensitivity and specificity when pooling data with nested structure for statistical testing. Samples were independently drawn for four subjects, $s = 1, \dots, 4$, and two variables, X and Y , according to $x_{n,s} \sim \mathcal{N}(\mu_s - 1, 4)$, $y_{n,s} \sim \mathcal{N}(\mu_s + 1, 4)$, where offsets μ_s were drawn independently for each subject from $\mathcal{N}(0, 15^2)$. A) Depiction of the means and standard errors for each subject. A significant difference between means is correctly identified for each subject, but not for the pooled data of all subjects (see lower panels). This is because of the substantial between-subject variance (see upper panels). B) Depiction of the data as a function of sample number (upper panel) and as scatter plots (lower panels). The common subject-specific offsets of X and Y cause strong correlation in the pooled data, which is not present in any individual subject, and may be considered spurious.

fect sizes $\hat{\theta}_s$ as group-level observations. In this approach, which is somewhat popular in the neuroimaging literature [22, 35], the null hypothesis $H_0 : \theta = \theta_0$ is tested based on the S subject-level effect sizes $\hat{\theta}_1, \dots, \hat{\theta}_S$, which are considered i.i.d.. The variance of the mean effect $\hat{\theta} = 1/S \sum_{i=1}^S \hat{\theta}_i$ is estimated as

$$\widehat{\text{Var}}(\hat{\theta}) = \frac{1}{S-1} \sum_{s=1}^S (\hat{\theta}_s - \hat{\theta})^2, \quad (32)$$

which is an unbiased estimate of $\text{Var}(\hat{\theta})$ even if variances σ_s^2 vary across subjects [32]. If the θ_s are normal distributed (for example, because they represent the means of normal distributed or many quantities), the test statistic

$$t = \frac{\hat{\theta} - \theta_0}{\sqrt{\widehat{\text{Var}}(\hat{\theta})}} \quad (33)$$

is t-distributed with $S - 1$ degrees of freedom. This is the standard one-sample t-test applied to the individual effect sizes $\theta_1, \dots, \theta_S$.

The naive summary-statistic approach is valid both under the fixed-effect and random-effects models [32]. Its statistical power is, however, limited due to two factors. First, it assigns equal importance to each subject. This is sub-optimal if subject-level variances σ_s^2 vary across subjects (for example,

because of different amounts of recorded data). In this case, a weighting scheme taking into account subject-level variances is optimal (see Section 2.6.2). Second, the approach does not make use of all the available data, as only the group level data is used to estimate the variance $\text{Var}(\hat{\theta})$ through Eq. (32). However, even if subject-level variances σ_s^2 are constant across subjects, it is beneficial to make use of their estimates (see Section 2.6.1).

Both issues are addressed by the sufficient-summary statistic approach described in the next section. An empirical comparison of the statistical power of both approaches on simulated data is provided in Section 3.

2.6. Sufficient-summary-statistic approach

If estimates of the variances $\text{Var}(\hat{\theta}_s)$ of the subject-level effect sizes $\hat{\theta}_s$, $s = \{1, \dots, S\}$ can be obtained, this gives rise to a more powerful summary-statistic approach compared to the naive approach outlined above. Assuming that the subject-level effect size estimates $\hat{\theta}_s$ are unbiased, any convex combination

$$\hat{\theta} := \frac{\sum_{s=1}^S \alpha_s \hat{\theta}_s}{\sum_{s=1}^S \alpha_s} \quad (34)$$

of the $\hat{\theta}_s$ with non-negative weights α_s , $s \in \{1, \dots, S\}$ is also unbiased (has expectation $\mathbb{E}(\hat{\theta}) = \theta$), as the denominator of Eq. (34) ensures that the weights sum to one. Importantly, with the exception of the coefficient of determination discussed in Section 2.2.4, all effect size measures discussed in this paper are unbiased estimators of the corresponding population effects. The variance of $\hat{\theta}$ defined in Eq. (34) is given by

$$\text{Var}(\hat{\theta}) = \frac{\sum_{s=1}^S \alpha_s^2 \text{Var}(\hat{\theta}_s)}{\left(\sum_{s=1}^S \alpha_s\right)^2}. \quad (35)$$

If the $\hat{\theta}_s$ are normal distributed (for example, because they represent the means of normal distributed or many quantities), the weighted mean $\hat{\theta}$ is also normal distributed. According to the central limit theorem, this is also approximately the case if the $\hat{\theta}_i$ are not normal distributed but the number of subjects S is large. In both cases, we can test the null hypothesis $H_0 : \theta = \theta_0$ using that the test statistic

$$z = \frac{\hat{\theta} - \theta_0}{\sqrt{\text{Var}(\hat{\theta})}} \quad (36)$$

is standard normal distributed.

The variances $\text{Var}(\hat{\theta}_s)$ typically need to be estimated, as the exact population values are unknown. As any estimate $\widehat{\text{Var}}(\hat{\theta})$ integrates information from all samples of all S subjects, it can be considered a fairly accurate estimate, justifying the use of a z-test even when we replace $\text{Var}(\hat{\theta})$ by its estimate $\widehat{\text{Var}}(\hat{\theta})$ in Eq. (36) [4, 6]. Sometimes, however, the more conservative t-distribution with $S - 1$ degrees of freedom is assumed for z [48, 25].

2.6.1. Equal-variance-weighting

The z-test introduced in Eq. (36) is valid regardless of the choice of the non-negative weights α_s , $s \in \{1, \dots, S\}$. One popular choice is to assign equal weights

$$\alpha_1 = \dots = \alpha_S = \frac{1}{S} \quad (37)$$

to all subjects, such that $\hat{\theta}$ becomes the arithmetic mean of the $\hat{\theta}_s$. This procedure is similar to the naive summary-statistic approach introduced in Section 2.5 in that both approaches assign equal importance to each subject-level effect size. However, it differs in the way the variance is estimated, and in terms of the distribution that is assumed for the test statistic. For the naive summary-statistic approach, variances are estimated through Eq. (32) using the S data points on the group-level only. The equal-variance-weighting approach instead uses the subject-level variances. That is, following Eq. (35): $\widehat{\text{Var}}(\hat{\theta}) = 1/S^2 \sum_{s=1}^S \widehat{\text{Var}}(\hat{\theta}_s)$.

If the individual $\widehat{\text{Var}}(\hat{\theta}_s)$ are unbiased, both methods yield an unbiased estimate of the variance $\text{Var}(\hat{\theta})$. But the variance of this variance estimate is typically smaller for the equal variance weighting approach, because it makes use of all the available data. This more accurate estimate means that the test statistic is approximately normal distributed rather than t-distributed with $S - 1$ degree of freedoms. This translates into a power gain, as illustrated in the simulation presented in Section 3. However, estimating the between-subject variance for a random-effects model is not straightforward, and also may introduce biases and variability (see Section 2.6.3).

2.6.2. Inverse-variance-weighting

Interestingly, the choice of equal weights is suboptimal in terms of obtaining a group-level effect size estimate $\hat{\theta}$ with minimal variance. It is generally desirable to minimize the variance of the weighted average, as unbiased estimators with smaller variance achieve a lower mean squared error (MSE), and lead to more powerful statistical tests. The minimum-variance estimate is obtained by weighting each subject-level effect size proportional to the inverse of its variance using weights

$$\alpha_s = \frac{1}{\text{Var}(\hat{\theta}_s)} \quad (38)$$

This result is consistent with the intuition that less precise θ_s should have a lower impact on the overall estimate than those that are estimated with high confidence. Inserting into Eq. (35), we obtain the optimal value

$$\text{Var}(\hat{\theta}) = \frac{1}{\sum_{s=1}^S 1/\text{Var}(\hat{\theta}_s)} = \frac{1}{\sum_{s=1}^S \alpha_s} \quad (39)$$

2.6.3. Estimation of between-subject variance

To perform inverse-variance-weighting under the random-effects model, the between-subjects variance σ_{rand}^2 needs to be estimated in order to obtain the total subject-wise variance $\text{Var}(\hat{\theta}_s) = \sigma_s^2 + \sigma_{\text{rand}}^2$. Several iterative and non-iterative alternative methods have been proposed [55, 16, 49]. A popular,

Algorithm 1 Inverse-variance-weighted sufficient-summary-statistic approach

Step 1: Within-subject analysis

for all Subjects $s = 1 \dots S$ **do**

Estimate effect size $\hat{\theta}_s$ and its variance $\hat{\sigma}_s^2$

end for

Step 2: Between-subject variance σ_{rand}^2

For random effects: use, e.g., Eq. (40)-(41)

For a fixed effect: $\hat{\sigma}_{\text{rand}}^2 \leftarrow 0$

Step 3: Population mean effect and variance

for all Subjects $s = 1 \dots S$ **do**

$\alpha_s \leftarrow 1/(\hat{\sigma}_s^2 + \hat{\sigma}_{\text{rand}}^2)$

end for

$\hat{\theta} \leftarrow \sum_{s=1}^S \alpha_s \hat{\theta}_s / \sum_{s=1}^S \alpha_s$

$\widehat{\text{Var}}(\hat{\theta}) \leftarrow 1 / \sum_{s=1}^S \alpha_s$

Step 4: Statistical inference ($H_0 : \theta = \theta_0$)

$z \leftarrow (\hat{\theta} - \theta_0) / \sqrt{\widehat{\text{Var}}(\hat{\theta})}$

z is approximately standard normal distributed

\Rightarrow Reject H_0 at 0.05 level if $|z| > 1.96$

relatively simple approach is the non-iterative procedure proposed by DerSimonian and Laird [9]. For a given estimate $\hat{\sigma}_s^2$ of the within-subject variances (which can be obtained using the procedures discussed in Section 2.2), and for fixed-effect quantities

$$\alpha_s^{FE} = \frac{1}{\hat{\sigma}_s^2}, \quad \hat{\theta}^{FE} = \frac{\sum_{s=1}^S \alpha_s^{FE} \hat{\theta}_s}{\sum_{s=1}^S \alpha_s^{FE}}, \quad (40)$$

the between-subject variance σ_{rand}^2 according to [9] is estimated as

$$\hat{\sigma}_{\text{rand}}^2 = \max \left\{ 0, \frac{\sum_{s=1}^S \alpha_s^{FE} (\hat{\theta}_s - \hat{\theta}^{FE})^2 - S + 1}{\sum_{s=1}^S \alpha_s^{FE} - \sum_{s=1}^S (\alpha_s^{FE})^2 / \sum_{s=1}^S \alpha_s^{FE}} \right\} \quad (41)$$

The truncation of the estimated variance to zero introduces a positive bias, that is, σ_{rand}^2 is over-estimated [39]. Moreover, the estimate may be quite variable for small sample sizes. As the inverse-variance-weighting approach does not take this variation into account, the resulting p-values can become too small when the number of subjects S is too small [5, 16]. Nevertheless, the DerSimonian and Laird approach is acceptable for a moderate to large number of subjects [25, 16], and is the default approach in many software routines in the meta-analysis community [49].

After $\hat{\sigma}_{\text{rand}}^2$ has been calculated, the random-effects quantities are finally computed as

$$\alpha_s^{RE} = \frac{1}{\hat{\sigma}_s^2 + \hat{\sigma}_{\text{rand}}^2}, \quad \hat{\theta}^{RE} = \frac{\sum_{s=1}^S \alpha_s^{RE} \hat{\theta}_s}{\sum_{s=1}^S \alpha_s^{RE}} \quad (42)$$

2.6.4. Algorithm

The sufficient-summary-statistic approach using inverse-variance-weighting is summarized in Algorithm 1. First, the subject-level effect sizes θ_s and the within-subject variances σ_s^2 , $s \in \{1, \dots, S\}$, are estimated based on the available subject-wise data samples. Second, the between-subject variance σ_{rand}^2 is estimated as outlined in Section 2.6.3 (unless a

fixed-effect model can reasonably be assumed). Third, the estimated population effect $\hat{\theta}$ is calculated as the average of the subjects effects, weighted by the inverse of their estimated variances as outlined in Section 2.6.2. This variance of a subject's estimated effect $\hat{\theta}_s$ around the population effect θ is the sum of the within-subject measurement error variance σ_s^2 and the between-subject variance σ_{rand}^2 (cf. Eq. (31)). Finally, the estimated mean effect is subjected to a z-test as introduced in Eq. (36).

Different effect sizes and their corresponding variances have been discussed in Section 2.2. With the exception of the Pearson correlation coefficient, these measures can be directly subjected to the inverse-variance-weighting approach. That is, $\hat{\theta}_s$ and $\hat{\sigma}_s^2$ for the mean difference are given in Eqs. (6) and (7), for the AUC in Eqs. (12) and (15), and for linear regression coefficients in Eqs. (23) and (24). As discussed in Section 2.2.4, it is, however, beneficial to transform correlation coefficients $\hat{\rho}_s$ into approximately normal distributed quantities with known variance prior to averaging across subjects. We can proceed with the application of the inverse-variance-weighting approach just as outlined before, treating the transforms $\zeta(\hat{\rho}_s)$ given in Eq. (18) rather than the $\hat{\rho}_s$ as effect sizes. The resulting population effect can be transformed back into a valid Pearson correlation using the inverse transformation described in Eq. (21).

2.7. Stouffer's method of combining p-values

A general approach for combining the results of multiple statistical tests is Stouffer's method [46, 51]. For a set of independent tests of null hypotheses $H_{0,1}, \dots, H_{0,S}$, Stouffer's method aims to determine whether all individual null hypotheses are jointly to be accepted or rejected, or, in other words, if the global null hypothesis $H_0 : (\forall s : H_{0,s} \text{ is true})$ is true. In general, the individual $H_{0,s}$ may not necessarily refer to the same effect size or even effect size measure, and the p-values for each individual hypothesis may be derived using different test procedures including non-parametric, bootstrap or permutation-based tests. In the present context of nested data, Stouffer's method can be used to test group-level null hypotheses in the fixed-effect setting, i.e., the absence of an effect in all S subjects of the studied population.

Denote with $H_{0,s} : \theta_s = \theta_0$ the null hypothesis that there is no effect in subject s , and with p_s the *one-tailed* p-value of an appropriate statistical test for $H_{0,s}$. If the null hypothesis is true, p_s is uniformly distributed between 0 and 1 (see [33] for an illustration). Therefore, the one-tailed p-values p_s can be converted into standard normal distributed z-scores using the transformation

$$z_s := F_{z_{0,1}}^{-1}(p_s), \quad (43)$$

where $F_{z_{0,1}}^{-1}$ denotes the inverse of the standard normal cumulative distribution function. For Gaussian-distributed subject-level effect sizes with known variance, this step can be carried out more directly using

$$z_s = \frac{\hat{\theta}_s - \theta_0}{\sqrt{\text{Var}(\hat{\theta}_s)}}. \quad (44)$$

The cumulative test statistic

$$z = \frac{1}{\sqrt{S}} \sum_{s=1}^S z_s \quad (45)$$

follows the standard normal distribution, which can be used to derive a p-value for the group-level H_0 .

Notice that Stouffer's method as outlined above is applied to *one-tailed* p-values. However, testing for the presence of an effect often requires a two-tailed test. In this case, it is important to take the direction of the effect in different subjects into account. We cannot simply combine two-tailed tests – a positive effect in one subject and a negative effect in another subject would be seen as evidence for an overall effect, even though they cancel each other out. However, the direction of the effect can be determined post-hoc. To this end, one-tailed p-values for the same direction are calculated for each subject and combined as outlined in Eqs. (43) and (45) into a group-level one-tailed p-value p_1 . The group-level two-tailed p-value is then obtained as $p_2 = 2 \cdot \min(p_1, 1 - p_1)$ (see Eq. (1)-(3)) [51].

3. Simulations

In the following, we present a set of simulations, in which we compare the statistical approaches reviewed above to test for a difference between two class-conditional means in artificial data. We consider two conditions \mathcal{X} and \mathcal{Y} with true means $\mu_{\mathcal{X}}$ and $\mu_{\mathcal{Y}}$ and class-conditional mean difference $d = \mu_{\mathcal{Y}} - \mu_{\mathcal{X}}$. We want to test the null hypothesis $H_0 : \mu_{\mathcal{X}} = \mu_{\mathcal{Y}}$ or, equivalently, $H_0 : d = 0$. The following scenarios are investigated: 1) The data are generated either within a fixed-effect or a random-effects model. 2) The data are generated from either a Gaussian or a non-Gaussian distribution. In each scenario, we compare the methods' abilities to reject the null hypothesis when we vary the true class-conditional mean difference d .

Data for $S = 5$ or $S = 20$ subjects $s, s \in \{1, \dots, S\}$, were generated as follows. First, subject-specific class-conditional mean differences d_s were sampled according to

$$d_s = d + \xi_s, \quad \xi_s \sim \mathcal{N}(0, \sigma_{\text{rand}}^2),$$

where σ_{rand}^2 is the between-subject variance. For the fixed-effect model, we set $\sigma_{\text{rand}} = 0$, while for the random-effects model, we set $\sigma_{\text{rand}} = 0.2$.

We then sampled $N_{s,\mathcal{X}}$ data points for condition \mathcal{X} and $N_{s,\mathcal{Y}}$ data points for condition \mathcal{Y} from Gaussian distributions with variance v_s^2 and class-conditional means $\mu_{s,\mathcal{X}}$ and $\mu_{s,\mathcal{Y}} = \mu_{s,\mathcal{X}} + d_s$, respectively. A separate set of samples was drawn from non-Gaussian F(2,5)-distributions adjusted to have the same class-conditional means and variance. The number of data points, $N_{s,\mathcal{X}}$ and $N_{s,\mathcal{Y}}$, the class-conditional means, $\mu_{s,\mathcal{X}}$ and $\mu_{s,\mathcal{Y}}$, and the variance, v_s^2 , were randomly drawn for each subject such that v_s is uniformly distributed between 0.5 and 2, $N_{s,\mathcal{X}}$ and $N_{s,\mathcal{Y}} \in \mathbb{N}$ are uniformly distributed between 50 and 80, and the true mean of class \mathcal{X} , $\mu_{s,\mathcal{X}}$, is uniformly distributed between -3 and 3. In each scenario, the true class-conditional mean difference, d , was varied across $\{0, 0.05, 0.1, 0.15, 0.2, 0.25, 0.3\}$.

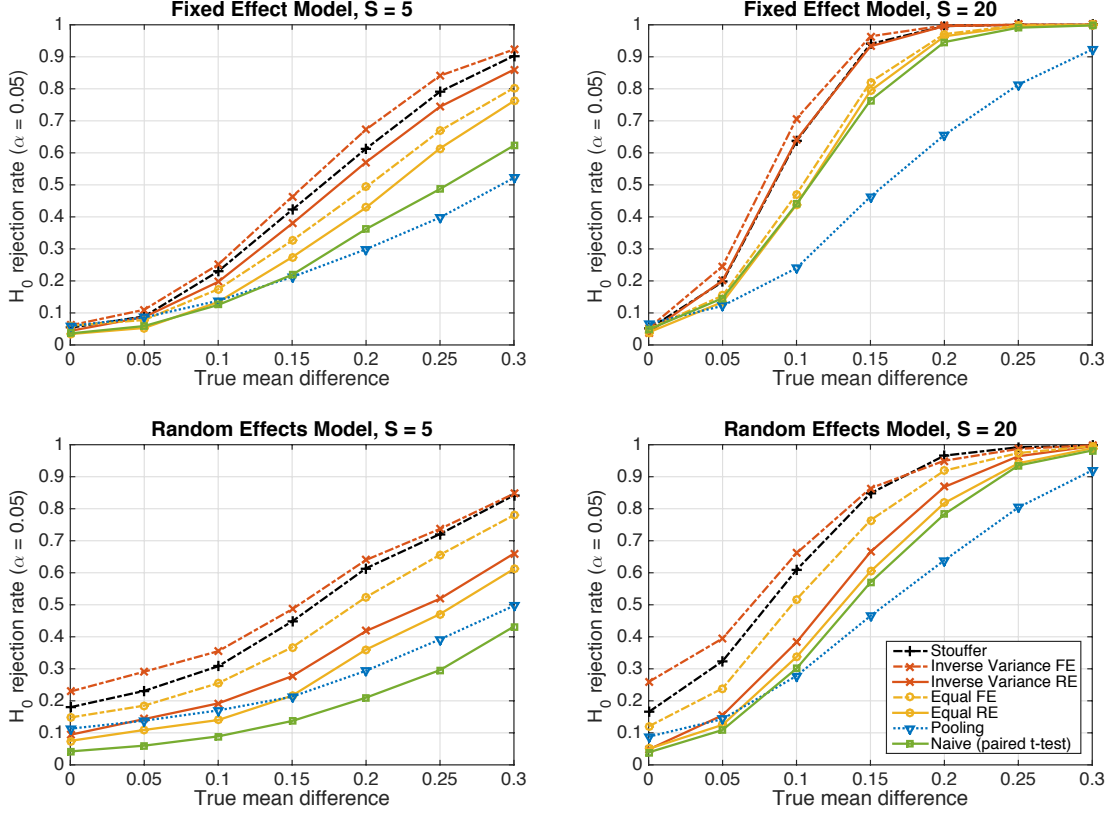


Figure 2: The probability of rejecting the null hypothesis $H_0 : d = 0$ as a function of the true mean difference d of Gaussian-distributed simulated data from $S=5$ resp. $S=20$ subjects. Top panel: data following a fixed-effect model. Bottom panel: data following a random-effects model.

All experiments were repeated 1000 times with different samples drawn from the same distributions. We report the H_0 rejection rate, which is the fraction of the test runs in which the null hypothesis is rejected. When the null hypothesis is true ($d = 0$), the H_0 rejection rate is identical to the error or false positive rate of the statistical tests under study. In the converse case, in which the null hypothesis is false ($d \neq 0$), the rejection rate determines the power of the test. All statistical tests were performed at significance level $\alpha = 0.05$. An ideal test would thus obtain a H_0 rejection rate of 0.05 when the null hypothesis is true ($d = 0$), and a rejection rate of 1 otherwise. The higher the H_0 rejection rate in the presence of an effect ($d \neq 0$), the higher is the power of a test. However, if the null hypothesis is true, a H_0 rejection rate greater than α indicates the occurrence of spurious findings beyond the acceptable α -level.

3.1. Simulation 1: Fixed effect vs. random effects

Figure 2 depicts the results achieved by the tested statistical procedures in the fixed-effect (top row) and random-effects (bottom row) scenarios for Gaussian-distributed data, using data from $S=5$ and $S=20$ subjects. The ‘pooling’ approach consists of pooling the samples of all subjects and performing one two-sample t-test (cf. Section 2.4). ‘Naive (paired t-test)’ refers the naive summary-statistic approach, in which each subject’s mean difference is treated as an observation for a group-level paired t-test (cf. Section 2.5). Four variants of the sufficient-summary-statistic approach are considered (cf.

Section 2.6). These variants differ in assuming either random effects (RE) or one fixed effect (FE), and in using either the inverse-variance-weighting scheme (Eq. (38)) or equal weights (Eq. (37)). ‘Stouffer’ finally refers to using Stouffer’s method to combine p-values obtained from subject-level two-sample t-tests (cf. Section 2.7). Note that all group-level tests are carried out two-tailed.

In line with our previous considerations, data pooling yielded very low power in the presence of an effect both under the fixed-effect and random-effects models. The highest power is achieved in both cases by the inverse-variance-weighted sufficient-summary-statistic approach, followed by Stouffer’s method, the sufficient-summary-statistic approach using equal weights, and the paired t-test.

Considerable differences are observed between the fixed-effect and random-effects settings. For data following the fixed-effect model, the fixed-effect variants of the sufficient-summary-statistic approach display only a negligible advantage over their random-effects counterparts, indicating that the latter succeed in estimating the between-subject variance to be zero. Moreover, in the case of equal class means, all approaches achieve a false positive rate close to the expected value of $\alpha = 0.05$.

The situation is different for data following a random-effects model. Here, the fixed-effect variants of the sufficient-summary-statistic approach as well as Stouffer’s method and the pooling approach display false positive rates that are between two and five times higher (26%) than what would be acceptable under the null hypothesis. This problem is sub-

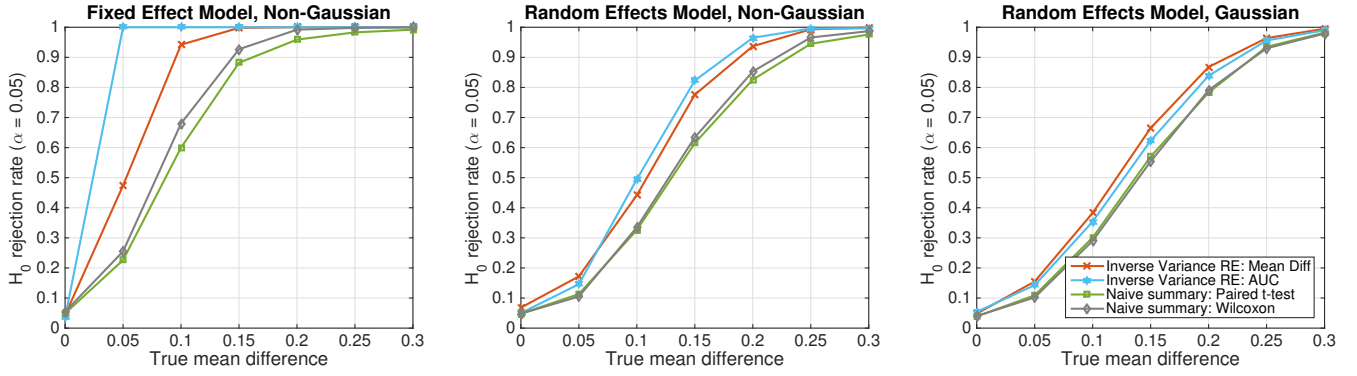


Figure 3: The probability of rejecting the null hypothesis $H_0 : d = 0$ as a function of the true mean difference d in simulated data of $S = 20$ subjects. (Left) Non-Gaussian data from a fixed effect model. (Middle) Non-Gaussian data from a random effects model. (Right) Gaussian data from a random effects model.

stantially alleviated by the random-effect variants of the sufficient-summary-statistic approach. Nevertheless, when data is only available from $S = 5$ subjects, the null hypothesis is still rejected too often (9% for inverse-variance-weighting). This is due to the variability in the estimate of the between-subject variance σ_{rand}^2 (cf. Section 2.6.3). When $S = 20$ subjects are available, the expected false positive rate of $\alpha = 0.05$ is achieved.

The naive summary-statistic approach (paired t-test of subject-wise means) achieves the expected false positive rate of 0.05 regardless of the number of subjects, and therefore represents a valid statistical test also in the random-effects setting.

3.2. Simulation 2: Gaussian vs. non-Gaussian

Figure 3 depicts the results of parametric and non-parametric statistical tests for simulated non-Gaussian-distributed data of $S = 20$ subjects following either the fixed-effect model (left panel) or the random-effects model (middle panel). For comparison, the results obtained on Gaussian-distributed data following a random-effects model are displayed in the right panel. Four different statistical tests are compared: 1) the random-effects inverse-variance-weighted sufficient-summary-statistic approach for the difference between class-conditional means, 2) the same test for the area under the non-parametric receiver-operating curve (AUC), 3) the naive summary-statistic approach in the form of a paired t-test between subject-wise means, and 4) its non-parametric equivalent, the Wilcoxon signed rank test. Note that for the naive summary-statistic approaches, the mean differences of each subject are treated as observations for a group-level paired t-test or Wilcoxon signed rank test, respectively.

The figure shows that, as for Gaussian-distributed data, the inverse-variance-weighted sufficient-summary-statistic approach achieves considerably higher statistical power than the corresponding naive summary-statistic approaches. Furthermore, non-parametric approaches achieve a higher power for non-Gaussian-distributed data than their parametric equivalents assuming Gaussian-distributed data. This difference is particularly pronounced for the better performing inverse-variance-weighted sufficient-summary-statistic approaches. The difference for the naive summary approaches is much

smaller, because subject-level averages tend to be more Gaussian according to the central limit theorem. In contrast, parametric approaches have only a very minor advantage over non-parametric ones for Gaussian-distributed data. Note further that, when the Gaussianity assumption of the parametric approaches is violated, spurious results can, in theory, not be ruled out. However, such effects are very small here.

4. Analysis of emergency-braking-related brain activity

We analyzed neuro- and myo-electrical activity of human participants during a simulated driving experiment. During the experiment, participants had the task to closely follow a computer-controlled lead vehicle. This lead vehicle would occasionally slow down abruptly, in which case the participant had to perform an emergency braking. The full study is described in [20]. Brain signals were acquired using 64 EEG electrodes (referenced to an electrode on the nose), while we here only report on the central EEG electrode Cz. Muscular activation of the lower right leg was acquired from two electromyographic (EMG) electrodes using a dipolar derivation. EEG and EMG Data were recorded from 18 participants in three blocks à 45 minutes. On average, clean data from 200 emergency situations were obtained from each participant (min: 123, max: 233). After filtering and sub-sampling to 100 Hz, the data were aligned relative to the onset of the braking of the lead vehicle as indicated by its brake light. For each time point relative to this stimulus, EEG and EMG measurements were contrasted with a sample of identical size that had been obtained from normal driving periods of each participant.

Figure 4 (top left) shows the deviation of EEG and EMG signals in emergency braking situations from signals obtained during normal driving periods as a function of time after stimulus. For each participant, a two-sample t-test was conducted (Eq. (6)). Assuming a random-effects model, the within-subject (i.e., within-participant) variance was estimated using Eq. (7), while the between-subject variance was estimated using Eq. (41). Results are presented in terms of the absolute value of the group-level z-score, which was computed using inverse-variance-weighting as described in Algorithm 1. It is apparent that the brain activity measured by EEG exhibits a significant amount of emergency-braking-related information

at an earlier point in time than the activity measured at the right leg muscle, but is superseded in terms of class separability by the EMG later on. This result reflects the decision-making process that is taking place in the brain prior to the execution of the physical braking movement.

The top right panel of Figure 4 depicts the same EEG time course in comparison to the curve obtained under the fixed-effect model. Compared to the RE model, the FE model leads to an inflation of z-scores starting 300 ms post-stimulus. Note that this is consistent with the result of Cochran’s Q-test for effect size heterogeneity [7] indicating non-zero between-subject variability after 200 ms post-stimulus ($p < 0.05$), but not before.

The bottom left panel of Figure 4 depicts the difference between the inverse-variance-weighted sufficient-summary-statistic approach and the naive summary-statistic approach implemented as a paired t-test for differences in the subject-wise means of the two conditions. As expected, the inverse-variance-weighted approach achieves a higher power than the naive approach by taking the subject-level variances into account.

Finally, the bottom right panel of Figure 4 depicts the difference between subject-level two-sample t-tests and non-parametric AUC tests according to Eqs. (12) and (15). Again, no substantial difference is found between the two, indicating that the raw samples are approximately normal distributed, justifying the use of a parametric t-test.

5. Discussion

In this paper we have provided a review of existing methods to assess the statistical significance of group-level effect sizes in data with nested structure. We demonstrated that simply pooling the data of all subjects is not a valid approach. In contrast, the naive summary-statistic approach of performing a paired t-test on subject-level effect sizes is valid. However, it has suboptimal statistical power. With the inverse-variance-weighted sufficient-summary-statistic approach and Stouffer’s method, we discussed two general strategies that combine the simplicity and low complexity of ‘naive’ approaches with higher statistical power by using prior knowledge about the distributions and variances of the subject-level effect sizes. The benefit of these two strategies over the ‘naive’ approaches was demonstrated in a set of simulations.

The simulations as well as the presented real-data analysis also highlighted the necessity to account for between-subject variances through a random-effects analysis. A failure to do so results in underestimated p-values and significant, but non-existing spurious effects. Stouffer’s method is a fixed-effects analysis, and thus provides a valid group-level test only if the assumption of zero between-subjects variance can be theoretically justified. In most practical cases, this is not the case [22, 12, 45, 41, 1]. We thus recommend the use of the inverse-variance-weighting approach when the number of subjects is modest and the subject-wise variances can be reliably estimated.

Importantly, while we here only considered data with two nesting levels, both Stouffer’s method and the inverse-

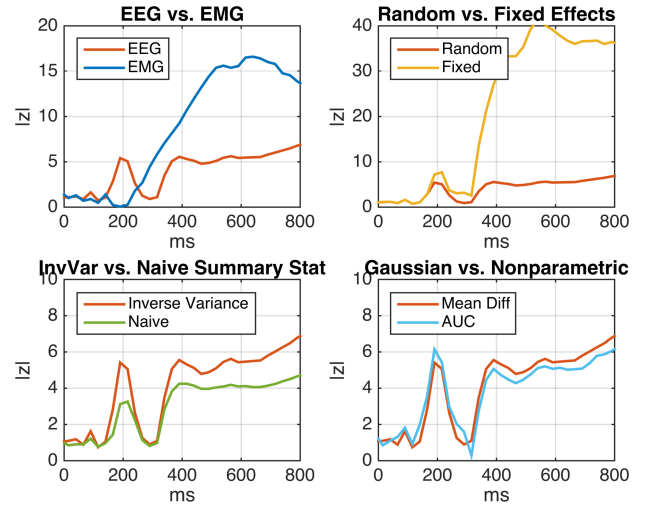


Figure 4: Analysis of event-related EEG (neural) and EMG (muscular) activity of $N = 18$ car drivers during simulated emergency braking. Shown is the z-scaled difference between the mean activity during emergency braking situations and the mean activity during normal driving periods as a function of time after the emergency-initiating situation. A) Comparison of EEG and EMG under the random-effects (RE) model using two-sample subject level t-tests and inverse-variance-weighting. EEG displays a significant class separation at an earlier time than EMG, reflecting the logical order of the underlying perceptual decision-making process. B) Comparison between the fixed-effect (FE) and RE models for EEG. The FE model displays inflated z-scores, indicating substantial but unaccounted between-subject variability. C) Comparison of the naive summary statistic approach and the sufficient-summary-statistic approach using inverse-variance-weighting for EEG (RE model). By taking the subject-level variances into account, the inverse-variance-weighted approach achieves a clearer statistical separation. D) Comparison between the two-sample t-test and the non-parametric Wilcoxon-Mann-Whitney test for a group-level area under the ROC curve (AUC) greater than chance-level (RE model, inverse-variance-weighting). Both tests lead to similar results, indicating that the distribution of samples is close to normal.

variance-weighted approach naturally extend to hierarchies with arbitrary numbers of levels. For example, p-values derived from individual subjects of a study, e.g., using Stouffer’s method, can again be combined at a higher level to test for consistent effects across multiple studies. In a similar way, group-level effects with variances derived from subject-level samples through Eq. (36) can be further combined into a higher-level average with known variance.

5.1. Alternative definitions of fixed and random effects

The notions of ‘fixed’ and ‘random’ effects are used differently in different branches of statistics. See, for example, [14] for a discussion of five different definitions of ‘fixed’ and ‘random’ effects in the statistical literature. In ANOVA, the factor levels of a ‘random effect’ are assumed to be randomly selected from a population, while the factor levels of a ‘fixed effect’ are chosen by the experimenter. In contrast to the definition of a ‘fixed effect’ used here (Eq. (27)), the effect sizes of a ‘fixed effect’ factor in ANOVA are allowed to differ across subjects.

Here we define a fixed effect (FE) to be constant across subjects, while a random effect (RE) is allowed to vary across subjects. The fundamental model underlying RE analysis is given by Eqs. (28) and (29), while the FE model is defined in Eq. (27). These definitions are used in the meta-analysis

literature [12, 4, 6], which contains most statistical discussion of between-subject variance estimators [9, 5, 41, 39].

In parts of the neuroimaging literature, a different interpretation of the fixed-effect model is predominant [35, 31]. Here,

$$\hat{\theta}_s = \theta_s + \epsilon_s, \quad (46)$$

where ϵ_s denotes the deviation of the subject's observed effect from the subject-specific true effect θ_s , which is not modeled as a random variable. In this view, the subjects are not randomly drawn from a population, but are 'fixed'. There is no overall population effect θ and the implicit null hypothesis behind the model is $H_0 : 1/S \sum_{s=1}^S \theta_s = \theta_0$. This yields an alternative interpretation of the same analysis: a fixed-effect analysis allows one to draw valid inference on the mean effect – but only for the specific mean of the observed subjects. Such an analysis would correspond to a case study, but a generalization to the population from which the subjects were drawn is not possible [35]. In contrast, the fixed-effect model Eq. (27) we assume throughout this paper allows such a generalization – but the assumption of a constant effect across subjects has to be theoretically justified.

5.2. Nested multiple linear models

Another approach to handle nested data are nested linear models (also called hierarchical linear models, multi-level models or mixed linear models). These models extend the multiple linear regression model discussed in Section 2.2.5 to deal with nested data. Following [23], this is done by introducing subject-specific regression coefficients $\beta_{k,s}$, $k \in \{0, \dots, K\}$, $s \in \{1, \dots, S\}$. The model for the n -th sample of subject s then reads

$$y_{n,s} = \beta_{0,s} + x_{n,1}\beta_{1,s} + \dots + x_{n,K}\beta_{K,s} + \epsilon_{n,s}. \quad (47)$$

The subject-specific coefficients are further expressed as

$$\beta_{k,s} = \gamma_{0,s} + \boldsymbol{\gamma}_s^\top \mathbf{z}_s + \epsilon_{k,s}, \quad (48)$$

where $\gamma_{0,s}$ is a subject-specific intercept, $\mathbf{z}_s = (z_{1,s}, \dots, z_{L,s})^\top$ models L known subject-resolved independent variables $z_{l,s}$, $\boldsymbol{\gamma}_s = (\gamma_{1,s}, \dots, \gamma_{L,s})^\top$ is a vector of corresponding coefficients modeling the influence of these variables on $\beta_{k,s}$, and $\epsilon_{k,s}$ is group-level zero-mean noise. In this complete form, all coefficients are subject-specific. We therefore speak of a random-effects nested linear model. It is also conceivable that only some of the coefficients are subject-specific, while others are shared between subjects. For example, in some applications it may be reasonable to model subject-specific intercepts $\beta_{0,s}$, but identical effects $\beta_{k,1} = \dots = \beta_{k,S} = \beta_k$ for all subjects. A resulting model would be called a mixed-effects nested linear model.

Nested linear models are very general and allow for more complex statistical analysis than the procedures for estimating and testing group-level effects discussed here. On the downside, the estimation of nested linear models is difficult because no closed-form solution exists in the likely case that the variances of the subject- and group-level noise terms are unknown. Fitting a nested linear model using iterative methods is time consuming when the number of subjects and/or

samples per subject is large, as all data of all subjects enter the same model. This is especially problematic when the number of models to be fitted is large, as, for example, in a mass-univariate fMRI context, where an individual model needs to be fitted for ten-thousands of brain voxels.

When only the group-level effect is of interest, the presented inverse-variance-weighting approach is the more practical and computationally favorable alternative. In this approach, regression coefficients $\hat{\beta}_{k,s}$ are estimated at the subject level, which bears the advantage that the global optimum for each subject can be found analytically in a computationally efficient manner. As the individual $\hat{\beta}_{k,s}$ are normal distributed with variance given in Eq. (24), they can then be combined using the inverse-variance-weighting scheme. This approach displays mathematical equivalence to a nested-linear model analysis when the covariances are known [2]. For these reasons, we here refrained from a deeper discussion of nested linear models. The interested reader is referred to, for example, [38, 54, 23].

5.3. Resampling and surrogate-data approaches

While the variances of the effect size measures discussed here can be derived analytically, this may not be the case in general. However, given sufficient data, the variance of the observed effect $\hat{\theta}$ can always be estimated through resampling procedures such as the bootstrap or the jackknife [10]. Assuming an approximately normal distribution for $\hat{\theta}$, the inverse-variance-weighting approach can hence be applied.

For some types of data such as time series, the subject-level i.i.d. assumption underlying most statistical procedures discussed here is violated. For such dependent samples, the variance of an observed effect $\hat{\theta}$ – be it analytically derived or obtained through a resampling procedure under the i.i.d. assumption – is underestimated. This problem can be addressed through sophisticated resampling techniques which accommodate dependent data structure. A detailed description of these techniques can be found, for example, in [26, 27].

For some types of analysis questions, it is not straightforward to determine the expected effect under the null hypothesis θ_0 . A potential remedy to this problem is the method of surrogate data, which was originally introduced in the context of identifying nonlinearity in time series [47] and is increasingly often applied to functional neuroimaging data in order to test for brain interactions (see [18] for a discussion). Surrogate data are artificial data that are generated by manipulating the original data in a way such that all crucial properties (including the dependency structure of the samples) are maintained except for the effect that is measured by θ . As such, surrogate data can provide an empirical distribution for $\hat{\theta}$ under the null hypothesis. This may be used to derive subject-level p-values, which can be subjected to Stouffer's method to test for population effects under the fixed-effect model.

5.4. Multivariate statistics

In the present paper we assumed that only a single effect is measured for each subject. However, massively multivariate data are common in many domains. This is particularly so in neuroimaging, where brain activity is typically measured at

hundreds or even thousands of locations in parallel. When (group) statistical inference is performed jointly for multiple measurement channels, the resulting group-level p-values must be corrected for multiple comparisons using, e.g., the false discovery rate [3].

Another way to perform inference for multivariate data is to use inherently multivariate effect size measures such as canonical correlations, coefficients of multivariate linear models, or more generally univariate effect size measures that are calculated on optimal linear combination of the measurement channels (e.g., [8, 19]). However, most multivariate statistics involve some sort of model fitting. If the number of data channels is high compared to the number of samples, overfitting may occur, and may bias the expected value of the effect under the null hypothesis. One way to avoid that bias by splitting the data into training and test parts, where the training set is used to fit the parameters of the multivariate model, while the actual statistical test is carried out on the test data using the predetermined model parameters [28].

5.5. Application in neuroimaging

While the methods summarized in this manuscript are applicable in any context, in which nested data occur, the examples provided here are all lent from neuroimaging. In parts of the neuroimaging literature, the use of suboptimal inference procedures such as the naive summary-statistic approach or data pooling is still widespread [32, 36]. Given the low signal-to-noise ratios and small sample regimes that are typical for neuroimaging studies, the loss of statistical power that is incurred by using these procedures is unfortunate. Nested linear models are frequently used in the analysis of fMRI data as implementations are provided by the major software packages. They are however less common outside the fMRI context. The sufficient-summary-statistic approach and Stouffer's method have occasionally been employed in neuroimaging contexts [34, 42, 20, 53]. The theoretical grounds on which they are applied have however not always been laid out in the same detail as here.

A distinction is made in the neuroimaging literature between so-called 'activation-like' and 'information-like' effect size measures. Allefeld et al. argue that measures that quantify the presence of an effect without a notion of directionality (that is, are 'information-like') cannot be subjected to a subsequent random-effects group-level analysis, because their domain is bounded from below by what would be expected under the null hypothesis of no effect [1]. Their arguments refer in particular to the practice of plugging single-subject classification accuracies into a group-level paired t-test. Because the true single-subject classification accuracies can never be below chance-level, the group-level null hypothesis being tested is the fixed-effect hypothesis of no effect in any subject. For the current investigation, this issue is, however, of minor importance, as, except for the coefficient of determination, all effect size measures discussed here are directional and therefore 'activation-like'.

6. Conclusion

In this paper, we have reviewed practical approaches to conduct statistical inference on group-level effects in nested data settings, and have demonstrated their properties on simulated and real data. With the inverse-variance-weighted sufficient-summary-statistic approach, we highlighted an approach that combines simplicity with excellent statistical properties. We have furthermore provided a practical guideline for using this approach in conjunction with some of the most popular measures of statistical effects.

Acknowledgements

SH was supported by a Marie Curie International Outgoing Fellowship (grant No. PIOF-GA- 2013-625991) within the 7th European Community Framework Programme.

References

- [1] Allefeld, C., Görgen, K., Haynes, J.-D., 2016. Valid population inference for information-based imaging: from the second-level t-test to prevalence inference. *NeuroImage* 141, 378–392.
- [2] Beckmann, C. F., Jenkinson, M., Smith, S. M., 2003. General multi-level linear modeling for group analysis in fMRI. *Neuroimage* 20 (2), 1052–1063.
- [3] Benjamini, Y., Hochberg, Y., 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 289–300.
- [4] Borenstein, M., Hedges, L. V., Higgins, J., Rothstein, H. R., 2009. *Introduction to Meta-Analysis*. Wiley Online Library.
- [5] Brockwell, S. E., Gordon, I. R., 2001. A comparison of statistical methods for meta-analysis. *Statistics in medicine* 20 (6), 825–840.
- [6] Card, N. A., 2011. *Applied meta-analysis for social science research*. Guilford Press.
- [7] Cochran, W. G., 1954. The combination of estimates from different experiments. *Biometrics* 10 (1), 101–129.
- [8] Dähne, S., Bießman, F., Samek, W., Haufe, S., Goltz, D., Gundlach, C., Villringer, A., Fazli, S., Müller, K.-R., 2015. Multivariate machine learning methods for fusing functional multimodal neuroimaging data. *Proceedings of the IEEE* 103 (9), 1507–1530.
- [9] DerSimonian, R., Laird, N., 1986. Meta-analysis in clinical trials. *Controlled clinical trials* 7 (3), 177–188.
- [10] Efron, B., Efron, B., 1982. *The jackknife, the bootstrap and other resampling plans*. Vol. 38. SIAM.
- [11] Fawcett, T., 2006. An introduction to ROC analysis. *Pattern recognition letters* 27 (8), 861–874.
- [12] Field, A. P., 2003. The problems in using fixed-effects models of meta-analysis on real-world data. *Understanding Statistics: Statistical Issues in Psychology, Education, and the Social Sciences* 2 (2), 105–124.
- [13] Fisher, R. A., 1915. Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. *Biometrika*, 507–521.
- [14] Gelman, A., 2005. Analysis of variance - why it is more important than ever. *The Annals of Statistics* 33 (1), 1–53.
- [15] Greiner, M., Pfeiffer, D., Smith, R., 2000. Principles and practical application of the receiver-operating characteristic analysis for diagnostic tests. *Preventive veterinary medicine* 45 (1), 23–41.
- [16] Guolo, A., Varin, C., 2015. Random-effects meta-analysis: the number of studies matters. *Statistical methods in medical research*, 0962280215583568.
- [17] Hanley, J. A., McNeil, B. J., 1982. The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology* 143 (1), 29–36.
- [18] Haufe, S., Ewald, A., 2016. A simulation framework for benchmarking EEG-based brain connectivity estimation methodologies. *Brain topography*, 1–18.

- [19] Haufe, S., Meinecke, F., Görgen, K., Dähne, S., Haynes, J.-D., Blankertz, B., Bießmann, F., 2014. On the interpretation of weight vectors of linear models in multivariate neuroimaging. *NeuroImage* 87, 96–110.
- [20] Haufe, S., Treder, M. S., Gugler, M. F., Sagebaum, M., Curio, G., Blankertz, B., 2011. EEG potentials predict upcoming emergency brakings during simulated driving. *Journal of neural engineering* 8 (5), 056001.
- [21] Hoerl, A. E., Kennard, R. W., 1970. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* 12 (1), 55–67.
- [22] Holmes, A., Friston, K., 1998. Generalisability, random effects and population inference. *Neuroimage* 7, S754.
- [23] Hox, J. J., Moerbeek, M., van de Schoot, R., 2010. Multilevel analysis: Techniques and applications. Routledge.
- [24] Hunter, J. E., Schmidt, F. L., 2000. Fixed effects vs. random effects meta-analysis models: implications for cumulative research knowledge. *International Journal of Selection and Assessment* 8 (4), 275–292.
- [25] Jackson, D., Bowden, J., Baker, R., 2010. How does the DerSimonian and Laird procedure for random effects meta-analysis compare with its more efficient but harder to compute counterparts? *Journal of Statistical Planning and Inference* 140 (4), 961–970.
- [26] Kreiss, J.-P., Paparoditis, E., 2011. Bootstrap methods for dependent data: A review. *Journal of the Korean Statistical Society* 40 (4), 357–378.
- [27] Lahiri, S. N., 2013. Resampling methods for dependent data. Springer.
- [28] Lemm, S., Blankertz, B., Dickhaus, T., Müller, K.-R., 2011. Introduction to machine learning for brain imaging. *NeuroImage* 56 (2), 387–399.
- [29] Mann, H. B., Whitney, D. R., 1947. On a test of whether one of two random variables is stochastically larger than the other. *The annals of mathematical statistics*, 50–60.
- [30] Mason, S. J., Graham, N., 2002. Areas beneath the relative operating characteristics (ROC) and relative operating levels (ROL) curves: Statistical significance and interpretation. *Quarterly Journal of the Royal Meteorological Society* 128 (584), 2145–2166.
- [31] Monti, M. M., 2011. Statistical analysis of fMRI time-series: a critical review of the GLM approach. *Frontiers in human neuroscience* 5, 28.
- [32] Mumford, J. A., Nichols, T., 2009. Simple group fMRI modeling and inference. *Neuroimage* 47 (4), 1469–1475.
- [33] Murdoch, D. J., Tsai, Y.-L., Adcock, J., 2008. P-values are random variables. *The American Statistician* 62 (3), 242–245.
- [34] Nolte, G., Ziehe, A., Nikulin, V. V., Schlögl, A., Krämer, N., Brismar, T., Müller, K.-R., 2008. Robustly estimating the flow direction of information in complex physical systems. *Physical review letters* 100 (23), 234101.
- [35] Penny, W., Holmes, A., 2007. Random effects analysis. In: Friston, K., Ashburner, J., Kiebel, S., Nichols, T., Penny, W. (Eds.), *Statistical Parametric Mapping*. Academic Press, London, Ch. 12, pp. 156–165.
- [36] Pernet, C. R., Chauveau, N., Gaspar, C., Rousselet, G. A., 2011. LIMO EEG: a toolbox for hierarchical linear modeling of electroencephalographic data. *Computational intelligence and neuroscience* 2011, 3.
- [37] Poline, J.-B., Brett, M., 2012. The general linear model and fMRI: does love last forever? *Neuroimage* 62 (2), 871–880.
- [38] Quené, H., Van den Bergh, H., 2004. On multi-level modeling of data from repeated measures designs: A tutorial. *Speech Communication* 43 (1), 103–121.
- [39] Rukhin, A. L., 2013. Estimating heterogeneity variance in meta-analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 75 (3), 451–469.
- [40] Ruxton, G. D., 2006. The unequal variance t-test is an underused alternative to Student’s t-test and the Mann–Whitney U test. *Behavioral Ecology* 17 (4), 688–690.
- [41] Schmidt, F. L., Oh, I.-S., Hayes, T. L., 2009. Fixed-versus random-effects models in meta-analysis: Model properties and an empirical comparison of differences in results. *British Journal of Mathematical and Statistical Psychology* 62 (1), 97–128.
- [42] Schubert, R., Haufe, S., Blankenburg, F., Villringer, A., Curio, G., 2009. Now you’ll feel it, now you won’t: EEG rhythms predict the effectiveness of perceptual masking. *Journal of cognitive neuroscience* 21 (12), 2407–2419.
- [43] Shorack, R. A., et al., 1966. Recursive generation of the distribution of the Mann-Whitney-Wilcoxon U-statistic under generalized Lehmann alternatives. *The Annals of Mathematical Statistics* 37 (1), 284–286.
- [44] Silver, N. C., Dunlap, W. P., 1987. Averaging correlation coefficients: should Fisher’s z transformation be used? *Journal of Applied Psychology* 72 (1), 146–148.
- [45] Stephan, K. E., Penny, W. D., Daunizeau, J., Moran, R. J., Friston, K. J., 2009. Bayesian model selection for group studies. *Neuroimage* 46 (4), 1004–1017.
- [46] Stouffer, S., Suchman, E., DeVinney, L., Star, S., Williams, R. J., 1949. *The American Soldier, Vol. 1: Adjustment during Army Life*. Princeton University Press, Princeton.
- [47] Theiler, J., Prichard, D., 1997. Using ‘surrogate surrogate data’ to calibrate the actual rate of false positives in tests for nonlinearity in time series. *Fields Inst Comm* 11, 99–113.
- [48] Thirion, B., Pinel, P., Mériaux, S., Roche, A., Dehaene, S., Poline, J.-B., 2007. Analysis of a large fMRI cohort: Statistical and methodological issues for group analyses. *Neuroimage* 35 (1), 105–120.
- [49] Veroniki, A. A., Jackson, D., Viechtbauer, W., Bender, R., Bowden, J., Knapp, G., Kuss, O., Higgins, J., Langan, D., Salanti, G., 2016. Methods to estimate the between-study variance and its uncertainty in meta-analysis. *Research synthesis methods* 7 (1), 55–79.
- [50] Welch, B. L., 1947. The generalization of Student’s problem when several different population variances are involved. *Biometrika* 34 (1/2), 28–35.
- [51] Whitlock, M. C., 2005. Combining probability from independent tests: the weighted Z-method is superior to Fisher’s approach. *Journal of Evolutionary Biology* 18 (5), 1368–1373.
- [52] Wilcoxon, F., Dec. 1945. Individual comparisons by ranking methods. *Biometrics Bulletin* 1 (6), 80–83.
- [53] Winkler, I., Haufe, S., Porbadnigk, A. K., Müller, K.-R., Dähne, S., 2015. Identifying Granger causal relationships between neural power dynamics and variables of interest. *NeuroImage* 111, 489–504.
- [54] Woltman, H., Feldstain, A., MacKay, J. C., Rocchi, M., 2012. An introduction to hierarchical linear modeling. *Tutorials in Quantitative Methods for Psychology* 8 (1), 52–69.
- [55] Worsley, K. J., Liao, C., Aston, J., Petre, V., Duncan, G., Morales, F., Evans, A., 2002. A general statistical analysis for fMRI data. *Neuroimage* 15 (1), 1–15.